DOCUMENT RESUME

ED 231 879 TM 830 487

AUTHOR Meskauskas, John A.
TITLE Standard-Setting: State of the Art, Future Prospects.
PUB DATE Apr 83
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec. April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Information Analyses (070)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Evaluation Criteria; *Evaluation Methods; *Research Methodology; Research Needs; *Research Problems; *Standards; *Summative Evaluation
IDENTIFIERS *Standard Setting

ABSTRACT

This paper presents the author's responses to nine questions addressed to the standard-setting symposium participants. A central theme in this assessment is that one cannot clearly demonstrate, a priori, that either a normative or a content-referenced approach is superior. Most methods used to date have been insufficiently grounded in psychological theory. Others have been inconsistent in the application of the theoretical base implicit to the method to the standard-setting problem. The concept of the minimally qualified candidate is discussed and found wanting from a theoretical perspective. It is proposed that standards be set in relation to the content being assessed. Characteristics of improved methods which would be firmly based in psychological theory are discussed. Two new methods for setting standards are suggested. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED231879

PRESENTATION FOR A SYMPOSIUM ENTITLED

STANDARD-SETTING: STATE OF THE ART, FUTURE PROSPECTS

BY:
JOHN A. MESKAUSKAS, PHD

Presented at the annual meeting of the
American Educational Research Association
Montreal, Canada

April, 1983

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. A. Meskauskas

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 487

Before addressing the various questions posed in this symposium, I wish to make a disclaimer. The perspective I offer has been formed from experiences with standard-setting problems in health certification and licensure situations. Therefore I will restrict my comments to this type of situation: summative examinations covering broad content domains. I make no claim that these views apply to the entire range of standard-setting situations.

QUESTION 1- SHOULD WE USE NORMATIVE OR CONTENT-REFERENCED STANDARDS?

For purposes of discussion, three, rather than two, categories of approaches to standard-setting may be distinguished. These are norm-referenced standards (NRS), content-referenced standards (CRS) and decision-theoretic standards (DTS). (For an excellent discussion of the first two types, see Livingston and Zieky (1982); for a representative discussion of DTS, see Huynh (1976, 1982) and Swaminathan, Hambleton and Algina (1975). DTS represent the newest development but they are also the least-used. I don't consider DTS methods viable for certification standard-setting, for reasons to be discussed, and will therefore concentrate on NRS and CRS.

Commonly, standards are set either in terms of the performance of a group (NRS) or in terms of the performance to be expected on a particular examination (CRS). The use of NRS implies that it is expected that candidate groups are equivalent over time and thus permits less concern with variation in the characteristics of any particular examination and over time. The use of CRS implies that standards are set in constant relationship to a content domain. Such standards are based exclusively upon the content of a particular examination and thus variation in the ability of the individuals undertaking that examination is irrelevant.

In an ideal world, CRS methods would appear to be preferable. Considering the state of development of standard-setting methods, I do not feel that a clear choice can be made, however. Two key issues in the choice of approach, beyond those raised in discussion of succeeding questions, include the mechanics of the standard-setting process and the degree to which the assumptions of a particular approach can be shown to hold in a particular application. NRS has an advantage of efficiency. Once the original decision is made, tradition makes subsequent standard-setting decisions very simple. The initial decision, however, is likely to be disquietingly arbitrary- based on a rather superficial study of the applicant characteristics, in many cases. As to the notion of stability of NRS norming groups over time, I am aware of only one study that has been undertaken to assess the extent of stability of such groups and the implications of any observed non-stability on standards (NBME, 1982). More such studies are needed in the medical specialties. In medicine, changes of the applicant pool, changing trends in specialization, flow and ebb of foreign medical graduates, etc., should make us very cautious about this assumption. The recent change in standard-setting policy of the NBME, (which deals with the largest and hence putatively most stable) group, is an important development.

CRS methods are relatively time-consuming to implement. Significant resources must be allocated to the study phase. Over successive administrations of an instrument, as new questions are written, either a new standard-setting study or an equating procedure must be carried out. As to assumptions, the key

assumption of CRS is that a standard can be specified in constant relationship to a content domain. The meaning and implications of this are not entirely clear in fields characterized by rapid change. If one were to plot the percent of professionals knowing a particular piece of information against the number of months elapsed from first publication, one would probably get a curve that grows rapidly in the first year or two, peaks, then, as the information becomes outdated, declines again. We can expect that the standard judges set for a test question assessing that information would be systematically related to such a dissemination curve.

Therefore, because of the present state of development of standard-setting methods, the best answer to the dilemma of which method to use is to utilize both NRS and CRS information in making the cutting-score decision. The situation is not yet clear enough to make discrete choices.

QUESTION 2- DIFFERENT STANDARD-SETTING METHODS YIELD DIFFERENT RESULTS. WHAT IS YOUR REACTION TO THIS FINDING?

Following the lead of Andrew and Hecht (1976), increasing numbers of studies have investigated the question of whether there are systematic differences in results associated with the choice of standard-setting method: Brennan & Lockwood (1980), Harasym (1981), Koffler (1980), Rock, Davis, & Werts (1980), Skakun & Kling (1980). All have found that this is the case. When I considered CRS methods (Angoff, Ebel, Nedelsky, etc.) as means for setting "absolute" standards, I was appalled. Since I now think of these methods as different means of setting content-referenced standards, I find these findings less troubling. Given the numerous situations in science where method influences results, why should we not expect that different standard-setting methods will yield different results?

The finding of strong method differences implies that they invoke very different processes and therefore that we should go beyond investigating method differences to investigations of the entire process within which the methods are imbedded. One neglected aspect is that of the issue of differences of examiner standards. We can take as well-established, (from studies with oral examinations, primarily), that substantial, stable, individual differences in examiner standards exist. Thus if we have sizeable variance associated with individual differences in the standards of judges, how does this variance relate to the variance associated with the choice of standard-setting method and, above all, to examinee variance? In some of my work with medical specialists, the variance due to examiner standard variability was large in relation to examinee variance. It may be that the choice of judges has as large an impact on the standard as does the choice of method. I would hope that those who have done these studies in the past might re-analyze their data to try to parse out the effects of individual differences and to look for interaction effects, particularly. Likewise, those considering studies in this area should plan to study this in a systematic way.

REACTION OF CLIENTS

In my personal experience, clients have been looking for "absolute" standards. When they become apprised of the fact that CRS methods are also highly judgmental, they have become disillusioned. NRS standards also involve judgment and the difference between the philosophical attractiveness of the two types narrows when one considers the available data.

DOES THIS FINDING PRESENT A PROBLEM FOR THE APPLICATION OF VARIOUS STANDARD-SETTING METHODS?

One might react to the finding that the situation is more complex than one had hoped by the type of nihilistic view espoused by Glass (1978) but a much more preferable perspective is taken by Shepard (1980). While I would preferthat the state of the art were more advanced, the practical realities are such that standards need to be set for various programs that are underway now. Thus we should not disdain from making reasoned decisions concerning choices of methods and applying those methods to real situations. As we do so, we should try to make the best decisions possible given the present circumstances and to simultaneously try to learn so that we improve decisions in the future. We should emphatically NOT disdain making decisions just because the methods are imperfect.

QUESTION 3: ASSESS THE THEORETICAL ADEQUACY OF THE GROUNDING OF VARIOUS METHODS OF STANDARD-SETTING IN PSYCHOLOGICAL AND/OR PSYCHOMETRIC THEORY

In considered this question, it very soon becomes clear that one basic problem of standard-setting is that there is too much focus on method and not enough on theory upon which to build that method. We have been re-discovering the inadequacies of "dust-bowl empiricism" all over again.

A consideration of the theoretical adequacy of any method must include the context in which it is used. There are a number of different situations in which we may wish to use normative normative standard-setting. We might wish to set a standard on an examination or perhaps an entire question pool and then maintain that standard over a number of administrations. In this application, excellenly reviewed by Shepard (1980), one might use either normative or content-referenced standards or a combination. Note, however, that doing this implies that the content domain of the field is not changing. I wish to examine what one might consider a "pure" form of NRS- where NRS is chosen explicitly because it is thought that the content domain is changing. This occurs in all professions, but is particularly obvious in medicine. Those who use norm-referenced standards in this way base their decision upon two key concepts- the notion of continuity of groups over time and a belief that distributions of reference groups (the subgroups upon which the standard-setting decision is made) will be unidimensional and normal in shape. There are no studies which test these assumptions in the literature. Thus all one can do is to look at the reasonability of these assumptions.

First, let us look at the notion that an educational training system produces groups of equivalent quality over long spans of time. An agency using this

type of standard measures a particular group of candidates, who constitute the output of an educational system, at one point in time. What factors might influence the degree of achievement of this group? We can identify as possible sources of variance the inherent talent of candidates (who self-select the particular profession at a point in time), individual differences in achievement of the candidates, the nature of the educational experiences (facilities, other resources), the preparation, motivation, and teaching skill of faculty, and the changing nature of a particular discipline. May we assume that, over a period of years, these forces balance out and produce equivalent results? When faced with the need to make an immediate, practical, decision concerning standards in an on-going program, perhaps we must. Over time, such an assumption seems highly improbable.

When an NRS standard is stated in terms of a Z-score or standard score derived from some group's performance, the agency responsible for the standard must be making the assumption that the examinee ability distribution is unidimensional and normal. (Were this assumption not made, even though it may not be made explicitly, it would not make any sense to compute a single composite standard score.) Note that the issue here is not so much the characteristics of the observed distribution as it is that of the theoretical distribution of scores. Is such an assumption reasonable? In most areas in which we certify and license canidates, probably not. Suppose we are working in an area where candidates have undergone successive steps of selection and self-selection. At the level of input to the training process, what type of theoretical ability distribution would this produce? On the output side, can we assume undifferentiated, uniformly trained, professionals? The goals of professional training are frequently couched in terms of mastery. It is not reasonable that this should generate groups with achievement distributions that are unidimensional and Gaussian in form. To be sure, this does not mean that we may not observe near-normal frequency distributions when we give a real examination. However, we should not accept such a finding as evidence that the underlying distributions are normal. It might simply mean that individuals differ in the number of examination-relevant areas they have mastered. For more discussion of this, see Bloom, Hastings, Madaus (1971).

From the above considerations it can be seen that the basic lack of NRS methods is that they, when practiced without a fundamental input-output understanding of the ways in which the characteristics of the educational system interact with the characteristics of individuals to produce the measured outcomes, are operating from largely untested assumptions.

Turning to a consideration of the theoretical foundation of CRS methods, one must note at the outset that such methods attempt a much more difficult feat than do NRS methods. The latter peg standards to more easily verifiable phenomena such as the performance of groups. CRS methods are linked to concepts that are much less straightforward and frequently internal to the standard-setter such as the notion of the minimally-qualified examinee or the relationship of a performance to a domain. These are difficult, and largely unexplored, cognitive problems.

Rather than examining the theoretical base of each method, key assumptions of various CRS methods will be addressed. (For the methods themselves, see

Livingston & Zieky, 1982; Meskauskas & Norcini, 1980; and Shepherd , 1980, for example.) We first turn to the notion that experts can provide us with accurate information regarding the expected examination-taking behavior of minimally-qualified examinees. This assumption is absolutely fundamental to several CRS methods. It seems a reasonable notion at first since we can visualize that such persons exist. To use such a group for standard-setting purposes, one has to be sure that the reason minimally-qualified persons are thought to be in the group is related to the domain the examination measures. Let us recall that Nedelsky (1954), who seems to be the originator of the notion of the minimally qualified examinee, applied the concept to setting a standard on a college course examination. Such an examination covers a very limited amount of material and it is not difficult to have a clear idea of the performance to be expected. Yet,many of the recent applications of CRS methods have been on summative examinations covering years of study. The larger and more complex the domain the more difficult it is to classify individuals as minimally-qualified on the basis of knowledge alone. In my experience with examinations in various areas of medicine, persons thought to be minimally qualified were viewed as such because some lacked knowledge, others had good knowledge but marginal problem-solving skills, others wavered in their ethics, and yet others were lacking in resolve to carry out their professional duties. To use methods that require a notion of a minimally-qualified examinee, it is essential that judges include individuals in such a group based solely on characteristics measured by the examination. In addition, judges must have had sufficient opportunity to observe a large enough number of minimally-qualified candidates so that the standard is not idiosyncratically linked to a small number of candidates.

Let us for the moment suppose that we are dealing with an area where experts can, in fact, intuit the existence of such a group with the type of precision that is needed. Can that mental concept be accurately translated into an expected behavior on a particular question or scale?

When a judge is asked to indicate how minimally-qualified persons would respond to a question, the response presumes that the group would respond in uniform ways. Yet we know that such individuals do not respond uniformly. On a particular item or sub-scale there is variation in performance by the members of such a group. Thus, what do we ask when we ask an expert to, in essence, predict how such a group would respond? What mental process and behavioral or group referents would be used? One can only speculate since there seem to be no studies of this. One possible approach might be for the expert to focus on a subset of the individuals in such a group, providing the expected response for this more uniform (but no longer representative) group. Alternatively, an expert may adopt a strategy that is different than that he is asked to follow. If this is done, and the responses appear to be consistent with the instructions, the experimenter might never know that this ocurred. Either approach is troublesome since the former is excesively linked to individuals while the latter is outside the design of the task. More generally, the notion of asking an expert to predict the performance of a more-or-less hypothetical group on a stimulus object which neither judge nor examinee have specifically experienced before is troublesome. It asks that the judge respond to the stimulus out of only the most general of contexts.

Some methods ask the judge to indicate the probability that a borderline candidate would correctly answer a question or a set of questions. This approach has the potential problem concerning the concept of a borderline candidate discussed above. In addition, it introduces the notion that experts can provide accurate probability estimates of the behavior of a group. Why should we assume that experts can do this? Experts are experts because of their knowledge of content, not because of their accuracy in providing probability estimates of the behavior of hypothetical candidates.

Other methods involve determinations concerning whether borderline candidates, who might not know the answer to a question, would in effect be allowed to choose one or more wrong responses. These methods require that a judge consider the stem of the question as well as the possible answers to it. Unlike other methods, this type of procedure requires the judge to process the question in a way that parallels that of the candidate and is consequently superior. While it invokes the questionable notion of the borderline candidate it poses a task which the judge can more directly relate to experience.

From a psychometric perspective, the lack of congruence between the method of scoring and the process by which the standard is set is troublesome. All the methods proposed heretofore develop a score by giving one point for the correct answer and zero points for a wrong answer. Thus it does not matter which right or wrong options are chosen, merely the overall number of them. Yet, during the standard-setting process, a substantial amount of information is gathered concerning the differences the judges perceive between options or questions. Some options are more important than others; some questions are more important than others. By ignoring this information, much is lost. We do not know whether there is, in fact, a close relationship between the judgments of the experts and the results of the scoring process. Further, in these times, when we are experiencing an explosion in computing power, it is inexcusable not to use all the information that is gathered.

SOME CHARACTERISTICS OF IMPROVED METHODS OF CONTENT-REFERENCED STANDARDS

Better methods of setting standards will result from a careful consideration of the standard-setting task within the context of the discipline being measured, the nature of the task posed for the experts, the characteristics of the examination, and the scoring system. An adequate theory will take all of these into consideration. Some of the characteristics that are essential to improved approaches are:

Design the experts' tasks so that judgments are provided in terms of the content of the field, not some hypothetical group. The content of a discipline is public and is therefore available to experts and candidates alike. Individual performance is not.

Utilize the capacity of experts to provide judgments concerning their field or concerning adequacy of performance. Focus on complex and valued characteristics, which the experts are uniquely qualified to judge.

Utilize what we have learned from the study of cognition to design the judges'

task.

Structure the task so that examinee behavior is being judged as directly as possible. Ask questions such as: "If the examinee chooses this response, what does it tell us concerning the examinee's capabilities or understanding?

Utilize an adequate number of judges to assure precision of the standard. Strive to structure tasks so that variability among judges is minimized.

Link the scoring system to the judgmental process.

TWO PROPOSED METHODS OF SETTING CONTENT-REFERENCED STANDARDS

Two methods, based on the foregoing considerations, will be proposed below. The first, let us call it method A, is constructed so that the experts' judgments parallel the task of the candidate and might be called a "micro-level" approach. The second is much broader. Influenced in its approach by the work of Brunswik (1952, 1955), Hammond (1966), and the more recent work of LaDuca (1980), method B asks judges to assess the quality of entire actual or simulated performances.

METHOD A

This method assumes several phases and is, of course, highly context-dependent. The first constitutes an extraction of the meaning that judges extract from the choosing of various options. The second explores the actual patterns of behavior of examinees. I will not attempt to prescribe the entire design this type of study would require. The unique steps that constitute this method are as follows:

1. Develop a classification system for the options (correct and incorrect) which incorporates both a qualitative or numeric differentiation among options and a standard of acceptability. This might take the form of: Best; acceptable; not acceptable; dangerous. The best approach to this step is one where the scale is closely linked to the subject matter in terms of the inference that may be drawn about a candidate who chooses a particular response.

2. Determine whether all questions are of equal importance or, if not, how many different categories of questions there are.

3. Develop scores, within question option classifications, for each class of option.

4. Conduct an item analysis on each option category. If unexpected patterns are found, consider whether the classification of those questions should be changed.

5. Study the relationship between unit scoring and the new system. In the example posed, the standard-setting question is one of how many "not acceptable" and "dangerous" questions should be allowed. This is judgmental since, even though it is tempting to assume that

those who chose dangerous options did so because they thought they were correct, there are many other reasons, including errors in the question phrasing, that might account for responses.

6. Study the relationship between the new system and traditional methods

7. When a decision is required, make the decision in terms of which method and level appears to be most reasonable.

METHOD B

This method focuses on extracting the judgment of experts regarding real or hypothetical performances of candidates. It is unconstrained, from the perspective of the expert, by considerations of the details of a scoring system. The steps are:

1. Determine the dimensions of performance that judges feel are assessed by the examination. This may be done in a number of differenct ways but should be done very carefully so that the end result reflects the judgments of the experts.

2. Administer the test and score in the usual binary manner

3. Choose a subset of a minimum of 30-50 people and the same number of questions. (These numbers will vary depending on the requirements of the specific application.) For each examinee, prepare multiple copies of this examination booklet, showing that examinee's responses to each of the questions.

3. Develop a behaviorally-based five to nine-point rating scale that has a pass/fail standard embedded within it.

3. Select a minimum of at least 5 judges and ask each judge to rate each of the candidates on the scale developed.

4. The ratings constitute a criterion measure to use to validate the items in the experimental scale as well as the scoring system. If the correlations are high, one can, by means of equating or item response theory, one can transfer the standard onto the entire examination.

   In several unpublished studies, Method B has yielded good agreement among judges and been an efficient, reliable, approach.

QUESTION 4- SHOULD STANDARDS BE VALIDATED? IF SO, HOW CAN THIS BE DONE?

Standards should be validated. As our discipline has developed, we have moved from the notion of validating instruments to that, under the rubric of generalizability theory, of validating inferences or interpretations (Cronbach

et. al., 1972). It is just a small step to suggest that we should primarily focus on validating a particulat type of infererence- the pass-fail decision. In carrying that out, one will automatically gather evidence regarding the validity of the instrument. The reverse does not necessarily hold.

As to the methods by which we might validate standards, it is clear that the major problem to overcome is not methodological but conceptual. Once we begin thinking about the notion of validating standards, it becomes clear that many of the methods widely used in psychometrics and experimental design would be applicable to this problem. Generalizability theory is certainly a useful conceptual framework for this type of work. Note that here one must distinguish from the perspective of generalizability theory, the logic of it, and the analysis of variance (ANOVA) methodology used within it. In summative examinations when examinees have been trained to mastery ANOVA methodology might prove misleading or, at best, not informative, since there may not be enough variance in performance for good ANOVA analysis

QUESTION 5- ROLES THAT ARE OR SHOULD BE PLAYED BY THE CLIENT, TECHNICAL CONSULTANT, CANDIDATES, AND OTHER ACTORS. WHAT IS THE PROPER ROLE OF THE PUBLIC?

QUESTION 6- SHOULD STANDARD-SETTING PROCESSES ATTEMPT TO FORMALLY INCORPORATE SOCIAL/POLITICAL CONSIDERATIONS INTO THE DECISION-MAKING PROCESS?

Upon reflection, these two questions are really part of one broad aspect- the social context of the standard-setting process. This is a very important but under-analyzed area. I will limit myself to a few short comments.

The client has the ultimate responsibility for making the standard-setting decision. In carrying that out, there is a natural tendency to make the decision very privately and quietly, to play things close to the vest. Less challenge will be aroused that way. The lawyers advise to disclose no more than absolutely necessary lest someone bring suit. Yet, in situations such as standard-setting, where the road to the best outcome is by no means clear, this approach closes off the client (or, if really closed off, the potential client) from input from others which will improve the quality of the process. I would hope that clients consider their decision in the most open way possible.

The role of the technical consultant is a difficult one. The major difficulty is that, although the consultant is frequently the person who brings important questions and (hopefully) incisive methodology to the client, the consultant's role does not include decision-making. This is probably for the best, but it is not a comfortable situation.

Candidates have an obvious interest in the nature of the decision. Given that, I have been continually amazed at the degree of ignorance among the candidates for medical specialty examinations concerning the tools, procedures, and standards that will be applied. In most instances, I believe that the overall quality of decisions would be improved if candidates began to ask constructive

questions of Boards.

As to the role of the public, we must first distinguish several different publics. There is the profession which directly or indirectly sets up some of its members to serve as a panel to confer a professional credential upon successful candidates. There is an obvious economic implication of the activities of certifying and credentialing bodies and this has attracted the interest of governmental agencies. Yet most of my personal experiences suggest that the vast majority of medical professionals, ignorant before the examination concerning the standards that will be applied, continue to remain ignorant even though successful. I continue to believe that the vast majority of professionals are deeply committed to quality of practice and their input would be helpful in establishing or refining the standards concerning those who will be certified in the first place. Constructive means are needed to allow this to happen.

The general public has an interest in who is certified. While I recognize the problems in finding people qualified to serve, certifying Boards should have input from the general public in their examination construction and standard-setting process. In addition, certifying bodies should publish information concerning their standard-setting process and results.

In general, I see the optimal model of standard-setting as an open one, with input from the many individuals and groups affected. This, of course, does not restrict the authority of the body responsible for setting the standards, but would hopefully help it carry out its responsibilities in an accoutable way.

QUESTION 7. ETHICAL RESPONSIBILITIES OF THE TECHNICAL CONSULTANT.

This is a very difficult question and broad answers are impossible. I believe that the technical consultant has ethical responsibilities in the standard-setting process. These responsibilities are to aid in making the most-informed decision possible. They also may be very difficult to carry out since not all clients are necessarily open to considerations beyond those that have already been brought to bear. There needs to be more awareness and consideration of this area. I would hope that a body like the AERA might take a lead in generating some discussion and guidelines in this area.

QUESTION 8. WHY HAVE DEVELOPMENTS COME SO SLOWLY? HOW DO YOU VIEW THE FUTURE OF STANDARD-SETTING?

Note: this Q needs organization. Bits and pieces which should be rearranged are below

It seems to me that standard-setting is one of the few areas in which there are more review articles than sound basic studies. This is one of a number of indications that we have fundamentally misjudged the problems. Our approaches have been treating standard-setting in a very narrow methodological way and have not been sound, psychologically. There has been insufficient basic work,

too many attempts to apply methods w hout a critical look at their explicit and implicit assumptions and the context of the problem.

Standard-setting is an intricate multiple-input process. Delopments have come slowly but I am encouraged by the developments of the last several years- the increased attention to the area from a variety of clients and researchers. Thus I see the future as brighter than the past but the going will be difficult.

QUESTION 9- WHAT ARE THE KEY SHORT-TERM AND LONG-TERM PROBLEMS THAT SHOULD BE ADDRESSED

As I am sure my previous comments have made clear, I strongly believe that a fruitful research approach in standard-setting is one that is multidisciplinary. Input from social psychology, sociology, practitioners in the field in which the examination is written, cognitive psychology, and perhaps other perspectives is needed in addition to the usual psychometric input. As to directions for research, first a general comment. Although the elegance of decision-theoretic models is very appealing, my belief is that this is not, except possibly for very special circumstances, a useful approach. The judges simply do not have any basis in experience upon which to base probability assessments. The student or practitioner groups they see are small, highly selected, and the judge may or may not have the opportunity to observe whether the critical behaviors occur. We are on much firmer ground when we base judgments on the content of the field, which is public (by means of publications, professional meetings, etc.) and hence much more verifiable.

NORMATIVE STANDARDS- RESEARCH PROBLEMS

Those who opt for NRS carry the responsibility to show that their assumptions are correct. That is, they should show that their approach does, in fact, lead to stable standards. Studies that need to be done in the short term include descriptive studies of the applicant pool and empirical validation of cutting score formulas. For the long-term, NRS advocates should study CRS.

CONTENT-REFERENCED STANDARDS- RESEARCH PROBLEMS

In the short term, task, environmental, and individual difference variables which influence judges' behaviors need to be studied. Also, instrument development on new methods is needed.

The studies above should be guided by the long-term perspective that standard-setting occurs in the context of several classes of variables: examinees, standard-setters, methods of examination and standard-setting, examination context, the content of the discipline, and the social/professional context within which the examination is taking place. All of these need to be considered in the long run. For more discussion of this, the interested reader may wish to see Meskauskas, 1980.

REFERENCES

Andrew, BJ, Hecht, JT. A preliminary investigation of two procedures for setting examination standards. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 36: 45-50

Bloom, BJ, Hastings, JT, Madaus, GF. HANDBOOK ON FORMATIVE AND SUMMATIVE EVALUATION OF STUDENT LEARNING. New York: McGraw-Hill, 1971

Brennan, RL & Lockwood, RE. A comparison of the Nedelsky and Angoff cutting score procedures using generizability theory. APPLIED PSYCHOLOGICAL MEASUREMENT, 1980, 4: 219-240

Brunswik, E. THE CONCEPTUAL FRAMEWORK OF PSYCHOLOGY. Chicago: Univ. of Chicago Press, 1952

Brunswik, E. Representative design and probabilistic theory in a functional psychology. PSYCHOLOGICAL REVIEW, 1955, 62:193-217

Cronbach, LJ, Gleser, GC, Nanda, H, Rajaratnam, N. THE DEPENDABILITY OF BEHAVIORAL MEASUREMENTS: THEORY OF GENERALIZABILITY FOR SCORES AND PROFILES. New York: John Wiley & Sons, 1972

Glass, GV. Standards and criteria JOURNAL OF EDUCATIONAL MEASUREMENT, 1978, 15:237-261

Gross, LJ Standards and criteria: a response to Glass' criticism of the Nedelsky technique. JOURNAL OF EDUCATIONAL MEASUREMENT, 1982, 19: 159-162

Hammond, KR Probabilistic functionalism: Egon Brunswik's integration of the history, theory, and method of Psychology. In KR Hammond (ed.) THE PSYCHOLOGY OF EGON BRUNSWIK. New York: Holt, Rinehart, Winston 1966

Harasym, PH. A comparison of the Nedelsky and Modified Angoff standard-setting procedure on evaluation outcome. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1981, 41: 725-734

Hambleton, RK, Swaminathan, H, Algina, J, & Coulson, DB. Criterion-referenced testing and measurement: A review of technical issues and developments. REVIEW OF EDUCATIONAL RESEARCH, 1978, 48: 1-47

Huynh, H. Statistical consideration of mastery scores. PSYCHOMETRIKA, 1976,

41: 65-78.

Huynh, H. A Bayesian procedure for mastery decisions based on multivariate normal test data. PSYCHOMETRIKA, 1982, 47: 309-320.

Koffler, SL. A comparison of approaches for setting proficiency standards. JOURNAL OF EDUCATIONAL MEASUREMENT, 17: 167-178

LaDuca, A. The structure of competence in health professions. EVALUATION AND THE HEALTH PROFESSIONS, 1980, 3: 253-288

Livingston, SA, and Zieky, MJ. PASSING SCORES. Princeton, NJ: Educational Testing Service, 1982

Meskauskas, JA. Future directions for research in standard-setting. Presented at: Board Evaluation procedures: Developing a research agenda. September, 1980. Evanston, Ill: American Board of Medical Specialties

Meskauskas, JA, Norcini, JJ. Standard-setting in written and interactive (oral) specialty certification examinations: issues, models, methods, challenges. EVALUATION AND THE HEALTH PROFESSIONS, 1980, 3: 321-360

NATIONAL BOARD EXAMINER. Philadelphia:National Board of Medical Examiners, summer 1982

Rock, DA, Davis, EL, & Werts, C. An empirical comparison of judgmental approaches to standard setting procedures. Research Report, June 1980, EDUCATIONAL TESTING SERVICE

Shepard, LA. Standard setting issues and methods. APPLIED PSYCHOLOGICAL MEASUREMENT, 1980, 4: 447-468

Skakun, EN & Kling, S. Comparability of methods for setting standards. JOURNAL OF EDUCATIONAL MEASUREMENT, 1980, 17: 229-235

Swaminathan, H, Hambleton, RK, & Algina, JA. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. JOURNAL OF EDUCATIONAL MEASUREMENT, 1975, 12: 87-98.

Zieky, MJ, Livingston, SA. A MANUAL FOR SETTING STANDARDS ON THE BASIC SKILLS ASSESSMENT TESTS. Princeton NJ: Educational Testing Service, 1977